Design and Package Technology Development of Face-to-Face Die Stacking as a Low Cost Alternative for 3D IC Integration

Zhe Li, Yuan Li, John Xie
Altera Corporation

Abstract

F2F stacking provides an alternative 3D packaging solution for multi-chip integration without use of TSV. High density interconnection can be achieved with direct Face-to-Face (F2F) stacking to enable high bandwidth die to die interface. Simplified stacking process and lower development cost make F2F stacking an attractive solution for cost sensitive applications. A comparative study of performance was performed on F2F stacked Field Programmable Gate Array (FPGA) die in a flip chip organic package. The paper first presents thermal analysis to address power density increase, hot spot and temperature variations in the F2F package. Next the paper focuses on electrical performance validation including both IO and power delivery analysis. With appropriate chip design and optimization, we demonstrate that F2F stacking induced thermal and electrical impacts can be controlled to meet speed and performance specs equivalent to 2D system. The manufacturing design rules have been optimized to meet yield requirements as well as ensuring product reliability.

Introduction

In recent years, 3D IC integration has attracted a great interest in semiconductor industry as a "More Than Moor" approach to break the silicon scaling trend [1]. Increasingly, traditional silicon process scaling alone can no long meet system performance, throughput and power requirements. 3D stacking enables more functional integration than single chip solution. Using through silicon via (TSV) to build 3D stacks of chips makes it possible to eliminate long run of interconnect lines in a 2D system. Power consumption is also reduced because of short interconnect path between active devices and lower IO drive strength to drive the short interconnects [2]. 2.5D and 3D packaging are being actively investigated. In order to achieve manufacturability readiness and cost effectiveness, the entire 3D fabrication sequence requires seamless integration from foundry to OSAT back-end process.

Though TSV technologies and 3D integration hold much promise, the industry also realized that the new technology will take time to get established in supply chain and drive volume production. While TSV and 3D technology development continues to improve, a whole variety of interim interconnect/packing solutions are being developed to address the bandwidth and scaling density demands. Direct Face-to-Face stacking through high density micro joints provides an alternative 3D packaging solution for multi-chip integration with low cost [3]. F2F stacking offers silicon grade interconnect density IO driver and reduced wiring parasitics for a high performance chip-to-chip interface. F2F packaging also has a time-to-market advantages compared to 3D IC which typically requires a longer lead time to implement design and longer cycle time due to the complexity of 3D manufacturing process.

F2F stacking provides additional benefits for compact form factors. There are some existing packaging solutions, such as side-by-side multi-chip package (MCP), package-on-package (POP) and package-in-package (PIP) platforms that feature footprint miniaturization and package height reduction. These packaging technologies allow for two or multi-chip to be bonded together. However, they do not offer the same degree of connectivity, scaling density, and bandwidth that F2F stacking can provide to meet the requirements of next generation products.

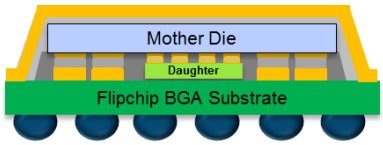

Figure 1. Face-to-Face stacked die on a BGA substrate as a low cost 3D IC solution without using TSV

This paper provides an overview of process flow and manufacturing activities for F2F stacking integration in FPGA applications. F2F stacking is a packaging technology designed to have two or more die assembled together. As shown in Figure 1, typically for two-die stacking, the mother die (top die) is larger than the daughter die (bottom die). Key feature of this technology is chip-on-chip bonding through fine pitch copper pillar micro-bumps. F2F stacking process flow is relatively simple. The critical steps are bumping and chip-on-chip assembly by Thermal Compression bonding with Non-Conductive Paste (TCNCP). TCNCP yield is sensitive to CoC alignment which depends on the planarity and topology of copper pillar micro-bumps [4]. Face-to-Face stacking is at early phase of technology development. The entire flow must be optimized to deliver overall yield and quality performance for manufacturing readiness and volume production

This paper is targeted at F2F package thermal and electrical design optimization. Package structural change from monolithic to stacked die presents unique challenges in the F2F package design. Thermal performance of F2F stacked die is discussed and compared to monolithic die in a flip-chip package. We also present a co-design approach in which RDL is employed together with die pinout and substrate routing to address the interconnect density and high speed channel bandwidth performance needs. High speed memory channel SI analysis is demonstrated to meet the channel loss and jitter requirements. Lastly, we discuss the PDN design and simulation methodology to accommodate F2F stacking impacts.

F2F Assembly Process Flow

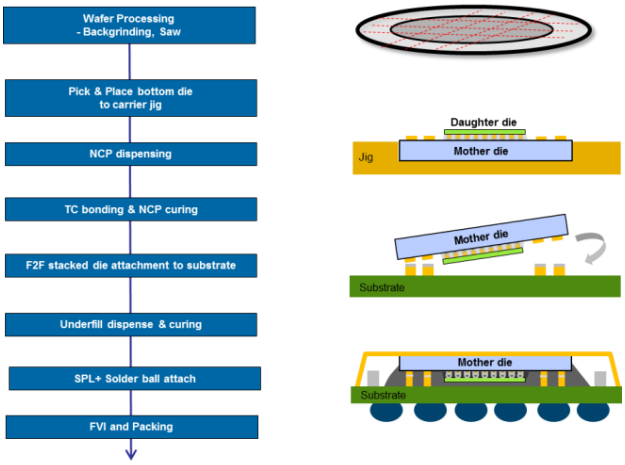

Figure 2. Process flow for Face-to-Face die stacking and substrate assembly

Face-to-Face assembly process is relatively simple. Figure 2 presents the process flow for F2F chip-on-chip bonding and then chip-on-substrate (CoCoS) assembly. After bumping is done on to the mother die wafer and daughter die wafer, back grinding and dicing are performed. The daughter die is then joined to the mother die through fine pitch copper pillar micro bumps. Thermal compression bonding with non-conductive paste (TCNCP) is employed for CoC assembly. As compared to the standard reflow process, TCNCP is chosen for CoC fine pitch assembly because of its accurate alignment control along with thermal compression to assist copper pillar joint formation. TC bonding consumes process time, but also provides process flexibility needed for F2F assembly to balance material stress for high yield and reliability. After completion with CoC assembly, the stacked dice are then flip-chip attached to the substrate through a mass reflow process. Capillary underfill is needed at this stage to ensure no air gaps form between die and substrate. The assembly is then followed by lid attach, ball attach, laser marking and test.

Package Structure

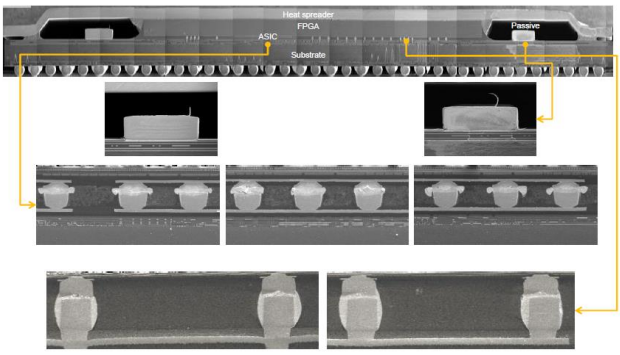

Figure 3. Cross section of test chip with FPGA die and daughter die face-to-face stacked with copper pillar joints

The F2F stacking process, interconnection material and package structures have been optimized for yield, reliability, electrical and thermal performance. Figure 3 shows the cross section and close-up view of the F2F assembled package and copper pillar joint structures. Two bump designs with different sizes and geometries are featured on the mother die. Fine pitch micro bumps are located in the CoC die pad region while the peripheral region is populated with large bumps in 185um pitch. An adequate amount of standoff clearance is needed so that die thickness variation and substrate warpage can be accommodated in CoS assembly. In F2F stacking, the daughter is thinned down to 75um. The CoC collapsed gap typically range from 25 to 35um depending on the bump designs. Total standoff height between die and substrate is on the order of 100 to 150um.

Cu Pillar Technology

Fine pitch copper pillar joint is the enabling technology for CoC assembly. For current face-to-face configuration, a taller copper pillar bump is created on the daughter die and a shorter height bump is formed on the larger mother die. The tall copper pillar bump is made up of 25um non-reflowable copper post with 15um SnAg solder cap. Table 1 shows more bumping options on both mother die and daughter die to make the fine pitch copper pillar joint. Pull-in and Pull-out designs are referring to the passivation opening relative to UBM size. If it is pull-out design, passivation opening is smaller than the UBM. Pull-in design has passivation opening larger than the UBM size. Both designs have been implemented and test out in DOE studies to reduce joint failures due to interconnect material stress. DOE result shows that pull-in design provides more bump contact area to silicon metal pads and thereby improves bump shear strength. Pull-out design has better protection on RDL layer and provides more design capability to reduce bump pitch to 40um and below.

Table 1. Fine pitch bumping technologies

	Daughter Die		Mother Die	
Die Thickness	75um		780um	
Bump Height	40um		3~5um	
Bump Composition	Cu Pillar/ LF Solder	Cu Pillar/ LF Solder	Ni/Au	Ni/Solder
Bump/PBO	Pull in	Pull Out	Pull In	Pull Out
Bump Structure				

Product Prototype Test Vehicle

To address the design and manufacturing challenges, we have designed a prototype F2F TV for FPGA integration with a high performance processor or memory. The F2F TV has a 0.18um 4mmx4mm CPLD chip mounted on a 40nm 17mmx22mm FPGA device through copper pillar joints in 50um bump pitch. The CoC stacked die is assembled on a 40mm organic substrate with 185um pitch copper bumps. The F2F TV contains an extensive set of structures for modeling and characterization of stacked die configuration and its impact to electrical and thermal performance. Characterization structures comprise of resistance, capacitance and S-parameter test structures for RDL, micro-bumps and substrate interconnects. F2F TV also contains varying critical area structures for yield and reliability characterization.

Thermal Analysis

In this work, the focus is on the comparison between F2F and monolithic packages. A F2F package and a monolithic single-die flip chip package are thermally analyzed using CFD (computationally fluid dynamics) models. In order to have fair assessment, both packages have the same package size, similar die size and the same total power consumption (26W). Both packages are modeled under the following conditions:

- JEDEC 2s2p PCB
- Ambient temperature 50 °C
- Al heat sink, same size as the package, plated fin, 20 fins, fin height 10 mm, thickness 1 mm, base thickness 1.5 mm
- Air flow rate 2 m/s

Table 2. T_j comparisons

	Tj of FPGA
F2F	101.1
Monolithic	98.8

The junction temperatures are listed in Table 2. The F2F package shows thermal performance similar to the monolithic package. With 26-W power, the difference in the junction temperature between the two packages is less than 3 °C. The higher Tj of F2F can be attributed to the smaller die size compared to the monolithic package. The temperature contours are shown in Figure 4.

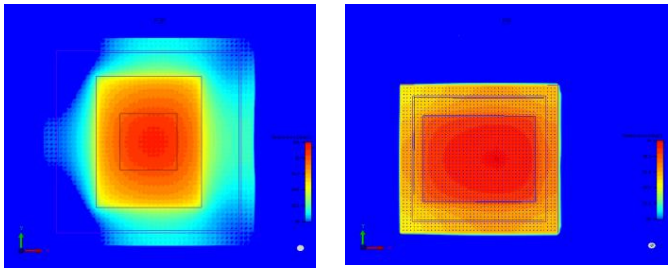

Figure 4. Die temperature contour on the monolithic (left) and F2F (right)

For the F2F package, under the conditions used, about 16W exits through the top and about 9.6 W exits through the bottom, or 60% out through the top and 37% the bottom

Thermal modeling has shown that F2F has comparable thermal performance to monolithic packages.

IO Electrical Performance

IO channel in F2F package has added 3 components, fine pitch copper pillar joints, redistribution layer (RDL) and tall copper joints between mother die and substrate. On substrates and below, IO channel routing is same as the conventional package. Figure 5 shows the RDL routing that connects IO channel from fine pitch bumps to large copper bumps. Due to stacked die spacing requirement, there are about 2~3mm routing length on RDL to fan out and get access to the substrate bumps.

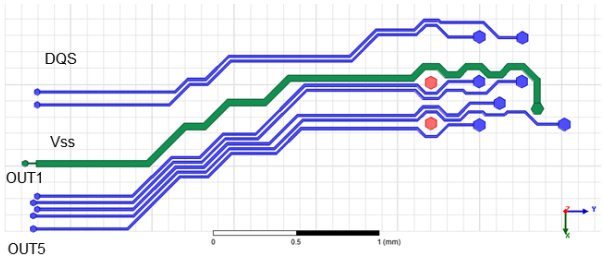

Figure 5. RDL routing for DDR channel in F2F stacked die package

DDR electrical validation was performed on data bus Read and Write operations for F2F stacked die package in comparison with conventional DDR channel design. The goal of this study was to validate the healthiness of DDR interface and to understand the performance difference between F2F stacked die and conventional packages. Package 3D and 2D models (Ansoft Q2D, Ansof HFSS) were carried out to simulate the DDR channel performance and then link simulation is used to run design of experiments (DOE) analysis to identify the weakest link in the design. The worst case bit pattern is generated to simulate the eye width and eye height. DOE analysis is a good way to assess the voltage and timing margins on DDR interface.

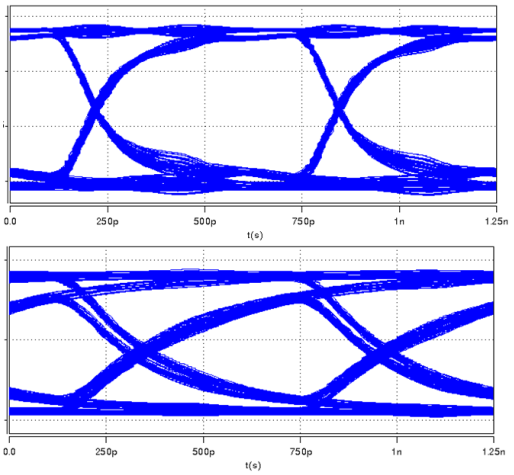

Figure 6. RDL impact to DDR channel eye opening (Top: DDR in flip-chip package, bottom: DDR in F2F package)

Link simulation of one DQ group data lines and DQS differential pair is carried out to capture the ISI, crosstalk and SSN impact from RDL routing and die-to-substrate interconnects. A snapshot of results of this study is shown in Figure 6 on DDR channel ISI due to RDL insertion loss. It is shown between F2F and conventional package using optimized routing and bump pattern, DDR performance degradation is about 8% in terms of voltage and timing margin. DDR performance degradation is mainly caused by RDL resistance. Another observation that can be made is that channel crosstalk in the F2F stacked die package are on par with conventional package. Channel crosstalk can be controlled by adding ground shielding to avoid potential coupling between active data lines

Power Integrity

Power integrity is one of the important issues in high speed system design. Power and current levels are expected to increase with corresponding to logic and I/O increment in F2F configuration. Some of the conventional wisdoms on power delivery network (PDN) analysis can still be applied to identify worst case power supply noise scenario.

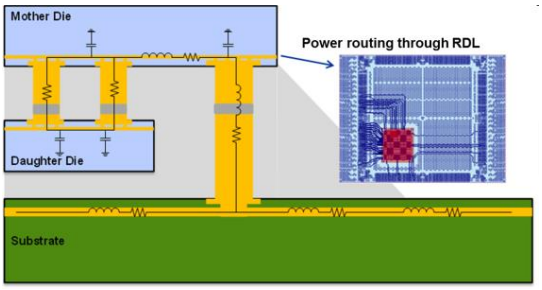

Figure 7. Power delivery network in F2F package

F2F package pose a new challenge on PDN design for its stacked die configurations and tightly coupled PDN system. The proposed face-to-face die stacking with two levels of copper pillar based joints and RDL interconnects is presented in Figure 7. The RDL has been discussed in the paper as a physical routing layer to link IO from the daughter die to C4 bumps and substrate. RDL also serves as the routing layer for power delivery. With limited routing space, metal stripping on RDL have to be balanced to meet both IO and power performance specifications. Power bus are connected from all four sides of the daughter die to minimize the potential DC resistance on power and ground.

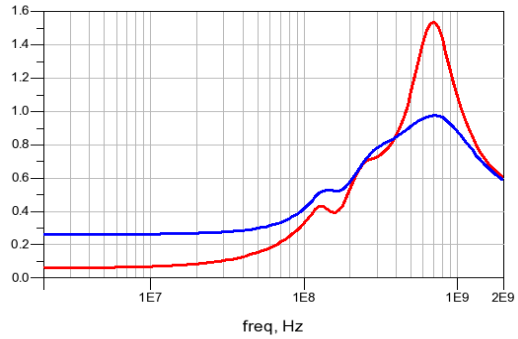

Figure 8. PDN impedance comparison between conventional package(red) and F2F package (blue)

In this study, interconnect parasitic impacts to PDN performance has been investigated in different packaging configurations. For power distribution, the preferred design is to reduce the interconnect resistance so that IR drop on power delivery network is small and DC voltage level at transistor power nodes is maintained. But it is not always feasible in practical designs, especially in F2F stacked die packages. RDL routing increases the DC resistance. Tall copper joint also adds resistance to the PDN. DC resistance needs to be managed in F2F package. Resistance impact to power stability is two folded. At high frequencies, resonance peak is reverse proportional to the PDN resistance. Usually the resonance peak occurs at a frequency determined by the on-die capacitance (ODC) and the inductance found in the package and PCB system. High resonance peaks give rise to large dynamic voltage droop in response to simultaneous switching activities. Previous work has demonstrated the issues caused by PDN impedance peaks. Methods have been developed to measure the PDN response to logic switching activities and correlate these with simulation results [5].

The parasitics in F2F stacked die package are extracted for comparison with conventional flip chip package and PDN impedance simulations are shown in Figure 8. The red line shows the PDN impedance in a conventional flip chip package design. The blue line shows the first pass power delivery performance of F2F stacking package as indicated in the test vehicle design. The simulation results confirm with our theoretical analysis. F2F stacked die package has a high DC resistance and low resonance peak than conventional packages. When considering the PDN design, balance the PDN performance from DC to high frequencies is the key for complete power delivery. In this study, RDL metallization, assignment of power and ground balls at several locations have been optimized to meet system power integrity requirements.

Conclusions

F2F stacking as a compelling and scalable package integration technology is demonstrated to address system bandwidth, low cost and small form factor needs. Such a new technology includes innovations in thin wafer handling, copper pillar micro-joint, die-to-die stacking as well as and die-to-package interconnections. Die-to-die interconnect is joined at 50µm pitch through TCNCP assembly. Key variables such as bonding force, bump co-planarity and warpage control have been optimized for F2F stacked die assembly. Test vehicles were designed and fabricated for interconnect reliabilities, thermal and electrical performance validations. Thermal and electrical modeling is carried out to verify F2F stacking impacts and make product design trade-offs.

Acknowledgments

The authors would like to thank Jemmy Sutanto, DongHe Kang, Michael Oh, Kwang-Seok Oh and their advanced product development team from Amkor for their support.

References

1. Knickerbocker, J.U., et al "3-D silicon intergration and silicon packaging technology using silicon through-vias," *IEEE Journal of Solid-State Circuits,* Vol. 41, No.8 (2006), page 1718- 1725
2. Hong Shi, Zhe Li, Yuan Li, Jenny Jiang, John Xie, "Performance Challenge and Opportunity of TSV Silicon Interposer in 3D System," *DesignCon* (2012)
3. Jemmy Sutanto, etc., "Development of Chip-on-Chip with Face to Face Technology as a Low Cost Alternative for 3D Packaging," *IEEE ECTC* (2013)
4. Zhe Li, SC Tan, YH Yew, PT Teh, MJ Lee, John Xie, "Flip-chip packages with Periphery Cu Pillar Bumps as Wire Bond Replacement- Design, Modeling & Characterization," *IMAPS* (2013)
5. Zhe Li, John Xie, Arif Rahman, "Development of an Optimized Power Delivery System for 3D IC Integration with TSV Silicon Interposer," *IEEE ECTC* (2012)